Cohesion in Automatically Created Summaries

Christian Smith, Henrik Danielsson, Arne Jönsson

Santa Anna IT Research Institute AB
Linköping, Sweden
`christian.smith@liu.se, henrik.danielsson@liu.se, arnjo@ida.liu.se`

Abstract

In this paper we present results from studies of cohesion in extraction based summaries. Cohesion is measured automatically through the amount of co-references in the text and how intact the text is after summarization. Four different ways of creating 100 word summaries are studied. The texts come from the DUC 2002 corpus of news paper texts. One interesting result is that a summary produced by a traditional vector space based summarizer is not less cohesive than a summary created by taking the most important sentences from the summarizer and the sentence before (at the same time removing the least important sentences), nor has the latter a lower content quality.

1. Introduction

An extraction based summary is created, by one way or the other, extracting the most important sentences from the original text. Cohesion and discourse relations play a vital role in understanding summaries (Louis et al., 2010). The goal is, of course, to create summaries that are readable in that they maintain cohesion while still maintaining a high coverage of content. Previous results have, however, shown that broken or erroneous anaphoric references is a problem in extraction based summarizers (Hassel, 2000; Kaspersson et al., 2012), as they are breaking the cohesion of the summarized text and in some cases even altering the meaning of the text, making them hard for readers to understand.

Pitler et al. (2010) attempted to develop and validate methods for automatic evaluation of linguistic quality in text summarization. Several human judges assessed automatically created summaries with regards to five topics; Grammaticality, Non-redundancy, Referential clarity, Focus and lastly Structure and Coherence. A set of indicators of linguistic quality was developed which then were used to rank summaries according to the five topics. Of these indicators, Continuity-based indicators performed the best in classifying summaries. Continuity-based indicators included Cohesive devices, coreference and cosine similarity.

Of the five topics, the topics of Referential clarity and Structure/Coherence seems to be the most relevant when dealing with extraction based single document summarization since 1) extraction based summarizers seldom have problems with grammaticality (Over et al., 2007), 2) single document summarizer is less likely to repeat information for redundancy and 3) for the same reason as 2) regarding Focus.

In this paper, we present results from studies on cohesion and summary content quality by comparing summaries created using various extraction techniques.

2. Method

In our experiments we created 100 word summaries of the 533 single document news paper texts from DUC 2002.

Four different types of techniques were used to create the summaries:

100FIRST extract the first 100 words, as a baseline, as for news texts the initial paragraphs normally includes the most important part of a text (Nenkova, 2005).

EVERY3 extract every third sentence,

COGSUM use a vector space based extraction summarizer,

PREVSUM use the summarizer and include also the sentence before the highest scored sentences and remove the lowest scored sentences, see Smith and Jönsson (2011).

The summaries were tagged for coreference using the Stanford CoreNLP package[1]. For our investigations we compare the summary to the original based on the following (for reference look at the short example text provided below):

> *The summer says bye.*
> *It has lasted rather long.*
> *The autumn is nigh.*

BROKEN if an anaphor has no antecedent (the first sentence in the example text has been removed).

INTACT if at least one antecedent to an anaphor is extracted (the last sentence in the example is removed).

PARTIAL if the anaphor is missing but the antecedent is in the summary (the second sentence in the example text has been removed).

REMOVED if a coreference chain is completely removed in the summary (The first and second sentences are removed).

Of these, BROKEN and INTACT are the most important as they affect cohesion the most. We also measure summary quality by gold standard comparisons using ROUGE-1.

The summarizer used in our investigations is a Random indexing based summarizer called COGSUM (Smith and Jönsson, 2011). The results however are valid for other vector space based summarization approaches e.g. HolSum (Hassel and Sjöbergh, 2007), Chatterjee and Mohan (2007) and Gong (2001).

---

[1] `nlp.stanford.edu/software/index.shtml`

Table 1: Summary quality with regards to content coverage

Summary	Content
EVERY3	0.36558
100FIRST	0.45925
COGSUM	0.39942
PREVSUM	0.38613

3. Results

Figure 1 depicts the results form our cohesion studies and Table 1 shows how the different summarizers performed.
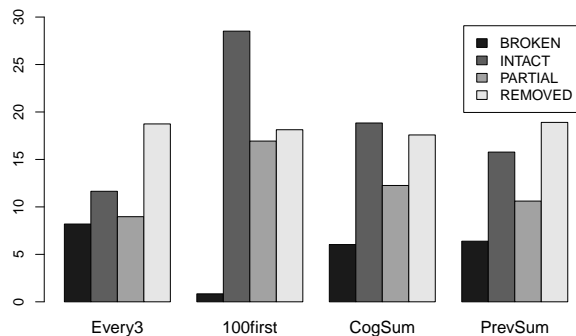

Figure 1: Cohesion measures on the four different summarizers.

Considering Content, 100FIRST is significantly better than all the other ($p < .05$). COGSUM and PREVSUM are significantly better than EVERY3 ($p < .05$).

Comparing the cohesion there are a variety of significances, for instance, for broken references the 100FIRST is significantly better than all the other ($p < .001$).

4. Conclusion

As expected the first 100 words of a text gives the best summary both in terms of cohesion but also summary content. This is true for short newspaper articles but probably not for longer texts from other genres. COGSUM, in its two versions, produce better summaries than taking every third sentence.

Taking the previous sentences in COGSUM doesn't affect the content quality and only slightly increases cohesion. This is interesting, as there is always a tradeoff between the amount of new information to include in the summary and the cohesion of the text. If some important sentences are disregarded and instead sentences that improve cohesion are included there is a risk that the summary will be less informative, but that was not the case in our studies using news texts.

Using news texts has its limitations, as also pointed out by Over et al. (2007), but this is where most current research is conducted, and is, thus, an interesting starting point. Further experiments with other text types may show that summarizers that consider cohesion give more readable summaries without significantly losing content.

5. References

Nilhadri Chatterjee and Shiwali Mohan. 2007. Extraction-based single-document summarization using random indexing. In *Proceedings of the 19th IEEE international Conference on Tools with Artificial intelligence – (ICTAI 2007)*, pages 448–455.

M Franzke, E Kintsch, D Caccamise, N Johnson, and S Dooley. 2005. Summary street®: Computer support for comprehension and writing. *Journal of Educational Computing Research*, 33(1):53–80.

Yihong Gong. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Martin Hassel and Jonas Sjöbergh. 2007. Widening the holsum search scope. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (Nodalida)*, Tartu, Estonia, May.

Martin Hassel. 2000. Pronominal resolution in automatic text summarisation. Master's thesis, Master thesis in Computer Science, Department of Computer and Systems Sciences (DSV), Stockholm University, Sweden.

Thomas Kaspersson, Christian Smith, Henrik Danielsson, and Arne Jönsson. 2012. This also affects the context - errors in extraction based summaries. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC), Istanbul, Turkey*.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Tokyo, Japan*, pages 147–156.

Ani Nenkova. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In Manuela M. Veloso and Subbarao Kambhampati, editors, *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 1436–1441. AAAI Press / The MIT Press.

Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context. *Information Processing & Management*, 43:1506–1520, Jan.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality inmulti-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden*, pages 544–554.

Christian Smith and Arne Jönsson. 2011. Enhancing extraction based summarization with outside word space. In *Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand*.